A System Engineering Approach to AI Security and Safety

Fariborz Farahmand, Georgia Institute of Technology

This article sheds light on the need for the computer science and the broader engineering communities to collaborate on taking a system engineering approach to artificial intelligence security and safety. I offer three recommendations to how to address this need.

As artificial intelligence (AI) affects our physical world, like driverless cars, robots, medical devices, and cyber physical systems, cybersecurity will enter those areas as well, and security and safety become increasingly entangled. That is, there is increasingly a potential for cybersecurity compromise to result in physical harm in the AI era, for example, manipulation of smart-building environments.

As such, the traditional divide-and-conquer approaches such as dividing the systems into separate parts, examining safety and security piece by piece, and combining the results no longer work. This creates the need, as explained by the National Academy of Engineers,[1] for a new understanding: "Engineers must recognize that a cybersecurity system's success depends on understanding the safety of the whole system, not merely protecting its individual parts." This article offers three recommendations to address this need.

IT IS ABOUT INTERACTION, NOT THE UNION

One aspect of AI safety research has focused on building robots and autonomous vehicles that do not collide. And AI security research has been focused on specific attacks, for example, input attacks that manipulate the data that is fed to the AI algorithm to affect the output of the system.

Digital Object Identifier 10.1109/MC.2023.3310219
Date of current version: 18 October 2023

To have safe and secure AI, we need to first understand complex interactions (for example, safety–security, software–hardware, human–automation) and develop realistic computational models that recognize these interactions and their interoperability issues. This requires a close collaboration between computer science and the broader engineering communities. So far, this has been limited to the collaboration between computer scientists and electrical and computer engineers, mainly their work on smart grids—the digital technologies that allow for two-way communication between the utility and its customers, and the sensing along the transmission lines. This unbalanced focus makes many of our critical physical infrastructures (for example, transportation and mechanical systems, and critical buildings which are already in poor physical condition with a C–rating from the American Society of Civil Engineers)[2] easy targets for cyber criminals.

EMBRACE THE ENGINEERING NOTION OF SAFETY WITH AN EYE TO THE SECURITY MEASURES

Engineers have a long history of research collaboration on reliability and safety. For example, electrical and computer engineers have built sensors and actuators that have significantly helped civil engineers to build intelligent infrastructures and monitor their safety and structural health. Two examples are the Golden Gate Bridge in the United States, for permanent seismic monitoring applications, and Moori Tower in Japan, for limiting deflections of the structure during typhoon winds.[3] But what are the reasons behind their successful interdisciplinary collaborations? Two important reasons are given.

First, engineers speak a common language: mathematics. Teaching engineering students from various disciplines, I have noted that mathematical explanation often steals their attention from the surrounding text that describes the problem and the solution, regardless of the title and the clarity of the text. They stay with this habit after graduation, even in dealing with nonengineers.

Second, engineers have developed modest and realistic expectations about safety. Engineers through experience have learned that instead of looking for perfect systems (for example, safe for all purposes); they will develop *good-enough* safe systems. That is, systems that would have increased safety are not worth the costs of reducing risk by restricting, or otherwise altering the service. Engineers use a probabilistic view of mathematical truth and social processes; very like the "social processes of mathematics to achieve successive approximations at understanding."[4] For example, to work with serviceability and safety issues, they define limit states: "conditions beyond which a structure or member becomes unfit for service and is judged either to be no longer useful for its intended function (serviceability limit state) or to be unsafe (strength limit state)."[4] Then, they formally define their problems in terms of limit-state functions by using probabilistic models of computation.

To present their system view on complex interactions, engineers and computer scientists must present their intuitive view of a system as a collection of interacting agents at different levels of abstractions.

Even engineering codes reflect the system and probabilistic view of engineers on safety. One example is safety of elevators which comfortably adapts the work of: The Structural Engineering Institute, The American Society of Civil Engineers, The American Society of Mechanical Engineers, and the Institute of Electrical and Electronic Engineers all working together to ensure user safety.[5]

This is quite different from the cybersecurity community, who over the past decades, has tried to develop perfectly reliable, dependable, trustworthy, safe systems, and use methods (for example, formal logic) to put a verified stamp on such systems.

USE CAUSAL ABSTRACTION FOR SYSTEM APPROACHES

Undoubtedly, AI has made an amazing progress in this century. It has already enabled us to do many things that we were not able to do before, for example, pattern recognition. But AI systems are not yet able to form humanlike abstractions in precise levels that we would need in taking a system approach. For example, deep learning models can give us ground-truth knowledge of the causal relationships between all their components, and can answer very abstract, high-level questions, for example, "what is?" questions. But they cannot answer "what if?" and "why?" questions.

To present their system view on complex interactions, engineers and computer scientists must present their intuitive view of a system as a collection of interacting agents at different levels of abstractions. To realize these abstractions, they also need tools that can help them with computational

TABLE 1. Pearl's three-level causal hierarchy.

Level	Typical activity	Typical questions
1. Association: $P(y\|x)$	Seeing (observing a certain phenomenon unfold)	What is? How would seeing *X* change my belief in *Y*?
2. Intervention: $P(y\|do(x), z)$	Doing (acting in the world to bring about some state of affairs)	What if? What if I do *X*?
3. Counterfactuals: $P(y_x\|x',y')$	Imagining (thinking about alternative ways the world could be)	Why? Was it *X* that caused *Y*? What if I had acted differently?

explanation of their views. To address these needs, I recommend causal abstraction, a general framework for explanation methods in AI. Its basic operation is intervention, which we can express using notation from Pearl.[6] Many popular behavioral and intervention-based explanation methods in AI can be explicitly understood as a special case of causal abstraction. Next, I provide a brief overview of Pearls' approach to causal abstraction in different levels, using *do*-operator and rules of *do*-calculus, and then present a simplified example.

Do-operator

Do-operator signifies that we are dealing with an intervention rather than a passive observation. *Do*-operator is different from the classical operators that we use in the standard language of probability.

- Observational $P(y|x)$ answers what is the distribution of *Y* given that we observe *X* variable takes value *x*.
- Interventional $P(y|do(x))$ answers what is the distribution of *Y* if we were to set the value of *X* to *x*.

Rules of *do*-calculus

Do-calculus, a calculus for probabilistic and causal reasoning (in Pearl's words, "machinery of causal calculus"[6]) is an axiomatic system for replacing probability formulas containing the *do*-operator with ordinary conditional probabilities. It uses three rules as follows:

- *Rule 1* helps us to ignore observations. It says when we observe a variable *W* that is irrelevant to *Y* (possibly conditional on other variables *Z*), then the probability distribution of *Y* will not change.
- *Rule 2* helps us to exchange actions with observations. It says if a set *Z* of variables blocks all backdoors from *X* to *Y*, that is, any path from *X* to *Y* that starts with an arrow pointing into *X*, then conditional on *Z*, *do* (*X*) is equivalent to observe (*X*).
- *Rule 3* helps us to ignore actions. It says we can remove *do* (*X*) from $P(y|do(x))$ in any case when there are no causal paths from *X* to *Y*. That is, if we do something that does not affect *Y*, then the probability distribution of *Y* will not change.

Three-level causal hierarchy

Table 1 outlines the three-level causal hierarchy, together with the characteristic questions that can be answered at each level:

- *Level 1*, association, invokes purely statistical relationships, defined by the naked data. Queries at this layer are placed at the bottom level in the hierarchy because they only present associations and not causal relations.
- *Level 2*, intervention ranks higher than association because it involves not just observing "what is" but changing what we observe.
- *Level 3*, counterfactuals is the highest level of the hierarchy because it subsumes interventional and associational questions.

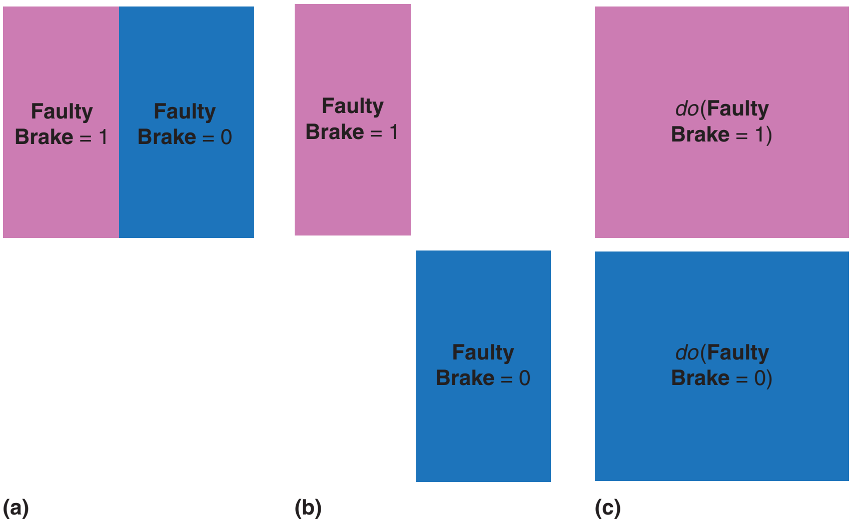

FIGURE 1. Difference between studying (a) observed data, (b) using conditioning, and (c) intervening, using *do*-operator.

A simplified example

"There is significant value in documenting and tracking AI failures in sufficient detail to understand their root causes, and to put processes and practices in place toward preventing similar problems in the future."[7] This example shows a simplified application of causal abstraction, at different levels, using *do*-operator, and rules of *do*-calculus in investigating the cause of an autonomous vehicle accident.

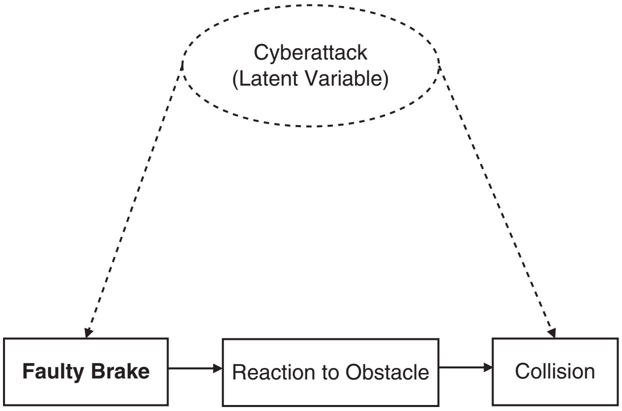

FIGURE 2. A graphical model representing the relationships among emergency braking, reaction to obstacle, collision, and cyberattack.

Assume a team of three experts, including a cybersecurity expert, a computer engineer, and a mechanical engineer are investigating the cause of an autonomous vehicle accident. The computer engineer and the mechanical engineer are considering an emergency braking system called **Faulty Brake**, as the probable cause of the accident. It was expected to react to the obstacle and perform an immediate and heavy braking maneuver, but it did not. They present some data that indicates some other cars that have used **Faulty Brake** were also involved in similar accidents.

Figure 1(a) shows their population (observed data) including two subpopulations, where **Faulty Brake** = 0 (shown in blue) is the subpopulation of the vehicles that did not use **Faulty Brake,** and **Faulty Brake** = 1 (shown in purple) is the subpopulation of the vehicles that used **Faulty Brake**. Figure 1(b) shows the process of conditioning on **Faulty Brake** = 1 and getting the purple subset on the top, or conditioning on **Faulty Brake** = 0 and getting the blue subset on the bottom. But the reality is that by conditioning they are just restricting the data to specific subsets of the data.

Now assume the cybersecurity expert informs our engineers of the possibility of cyberattacks shown in the dotted circle in Figure 2 that could influence both emergency braking functionality and the collision. But he does not provide any data. If our engineers can conduct randomized controlled experiments, known as the golden standard of statistics, they can address this concern. But unfortunately, many questions do not lend themselves to randomized controlled experiments. We cannot control the weather, so we cannot randomize the variables that affect wildfires. Similarly, our engineers cannot ask people to drive a car with a suspicious emergency braking system in traffic. Therefore, our engineers need to predict quantitatively the results of an intervention, that is, a Level-2 action, without actually performing it.

Figure 1(c), versus Figure 1(b), shows the Level-2 work of our engineers on intervening, versus conditioning. In Figure 1(c), engineers are not restricting the data to specific subsets. In *do* (**Faulty Brake** = 1) (in purple), they investigate what it would be like for vehicles in the population to use **Faulty Brake**. And similarly, for *do* (**Faulty Brake** = 0), they investigate what it would be for every vehicle in the population which did not use **Faulty Brake**.

To address the concern about the possibility of cyberattacks, our engineers need to use *do*-operator to assess *P*((Collision|*do*(**Faulty Brake**)). That is, the probability of collision given the autonomous vehicle had used **Faulty Brake**. Using probability axioms, this query can be expanded as shown in (1) at the bottom of the page.

Without available data to use *do*(**Faulty Brake**), our engineers can apply *do*-calculus rules to eliminate *do*-operators, and answer the query, using standard probability operators, as shown in (2) at the bottom of the page.

Do-calculus that was used in this example is just one example of the tools that can be used by cybersecurity and safety experts to facilitate collaboration with other experts, for example, engineers and policy and behavioral scientists. Other possible tools are probabilistic model checking (an extension of model checking techniques to probabilistic systems), and probabilistic programming (an extension of probabilistic graphical models, leveraging concepts from programming language research).

$$\sum_{\substack{\text{Reaction}\\ \text{to Obstacle}}} P\,(\,\text{Collision}\;\mid\; do\,(\textbf{Faulty brake}),\,\text{Reaction to Obstacle})\; P((\text{Reaction to Obstacle}\mid do\,(\textbf{Faulty brake}).\qquad (1)$$

$$\sum_{\substack{\text{Faulty}\\ \text{Brake'}}}\;\sum_{\substack{\text{Reaction}\\ \text{to Obstacle}}} P\,(\,\text{Collision}\;\mid \text{Reaction to Obstacle},\,\textbf{Faulty Brake'})\; P(\textbf{Faulty Brake'})\; P\,(\text{Reaction to Obstacle}\mid \textbf{Faulty Brake}).\qquad (2)$$

Integrating safety and security with AI is achievable and critical. My three recommendations: 1) It is about the interaction, not the union. 2) Embrace the engineering notion of safety with an eye to the security measures. 3) Use causal abstraction for system approaches are offered as one possible approach for success. Bring on the collaborators.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Award 2041788.

REFERENCES

1. National Academy of Engineering, “NAE grand challenges for engineering,” 2017. [Online]. Available: https://www.nae.edu/187212/NAE-Grand-Challenges-for-Engineering
2. ASCE, “Report card for America’s infrastructure,” 2021. [Online]. Available: https://infrastructurereportcard.org/
3. J. Lynch, H. Sohn, and M. Wang, *Sensor Technologies for Civil Infrastructures*, 2nd ed. USA: Elsevier, 2022.
4. R. A. De Millo, R. J. Lipton, and A. J. Perlis, “Social processes and proofs of theorems and programs,” *Commun. ACM*, vol. 22, no. 5, pp. 271–280, 1979, doi: 10.1145/359104.359106.
5. ASCE, “Minimum design loads and associated criteria for buildings and other structures,” 2022. [Online]. Available: https://ascelibrary.org/doi/book/10.1061/9780784415788
6. J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
7. F. Durso, M. S. Raunak, R. Kuhn, and R. Kacker, “Analyzing failures in artificial intelligent learning systems (FAILS),” in *Proc. IEEE 29th Annu. Softw. Technol. Conf. (STC)*, 2022, pp. 7–8, doi: 10.1109/STC55697.2022.00010.

FARIBORZ FARAHMAND is a research faculty member at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. He is a Senior Member of IEEE. Contact him at fariborz@ece.gatech.edu.

Digital Object Identifier 10.1109/MC.2023.3317994